# REDACT: PII Redaction and Privacy Protection with Ollama Integration

Mohamad Aasif J
Computer Science and Engineering
Sri Krishna College of Engineering and Technology
Coimbatore, India
aasif013010@gmail.com

Ms. Azhaguramyaa VR
Computer Science and Engineering Sri Krishna College of Engineering and Technology Coimbatore, India
vrazhaguramyaa@gmail.com

Mrinalini K
Computer Science and Engineering Sri Krishna College of Engineering and Technology Coimbatore, India
mrinalini.krishnadas@gmail.com

Madhu Priya R
Computer Science and Engineering Sri Krishna College of Engineering and Technology
Coimbatore, India
madhubhav4u@gmail.com

***Abstract:*** **Protection of Personally Identifiable Information (PII) in government documents is required for privacy and data security. Traditional redaction of PII is slow, time-consuming, and imprecise for bulk use. In this paper, an automated and secure PII redaction system with support for various document types, including PDFs and scanned images, is presented. Through the application of Optical Character Recognition (OCR) in extracting text and Large Language Models (LLMs) through Ollama in PII identification, our solution ensures accurate detection of sensitive information in various languages. Users have access to review and customize redaction rules in order to create flexibility and autonomy in the process. With the addition of advanced document processing techniques and visual redaction for image- based data, the solution suggested enhances security, ease of use, and conformance with data protection law.**

***Keywords: Personally Identifiable Information(PII), Large Language Models (LLMs), Optical Character Recognition (OCR), Ollama.***

## I. Introduction

The level of digitization of government and administrative processes has brought about widespread use of electronic documents for identification, transactions, and data exchange. These documents typically carry Personally Identifiable Information (PII), disclosure of which can result in identity theft, financial fraud, and privacy invasion [10].Regulatory frameworks such as the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and California Consumer Privacy Act (CCPA) necessitate stringent data protection measures in order to protect sensitive information [11]. However, with increasing digitalization, the possibility of unauthorized access and data breaches is more imminent, and efficient PII redaction systems become inevitable [12]. Legacy PII redaction solutions are severely constrained. Most are keyword-based or template-based, which are not able to process unstructured data, multilingual text, and handwritten text [3]. Moreover, current solutions do not have full facial redaction capability and offer users minimal control over redaction options [13]. These inefficiencies render them unsuitable for current large-scale document management [15].

To fill this void, we suggest an AI-based PII redaction solution for administrative and government reports such as Aadhar cards, driving licenses, voter IDs, and PAN cards.

We have handled various document types such as scanned documents, images, and PDFs so that the PII redaction is accurately and safely done [5]. With Optical Character Recognition (OCR)-driven text extraction [7][8], machine learning-driven intelligent detection of PII [6][12], and enhanced image processing driven by visual redaction [9], the solution delivers a scalable, privacy-accountable, and user-aware response. We present our redaction framework in this paper, compare it with other approaches, and illustrate how it improves digital document processing for security, usability, and compliance.

## II. Literature Review

PII protection has become a crucial issue following the emergence of e-government services, online platforms, and data breaches. Redaction processes based on manual or rule-based methods are ineffective with unstructured, handwritten, and multilingual data, proving to be inadequate for use in large-scale applications. They do not scale well and are not efficient with diversified document types and formats, resulting in real-world deployment constraints [10]. Specifically, the growing amount of sensitive information in various forms, such as images and scanned documents, requires more sophisticated methods that extend beyond traditional redaction techniques.

To address this, AI-based solutions, especially those founded on machine learning (ML), have advanced considerably in PII redaction automation. Machine learning algorithms can learn patterns in big data, allowing them to detect and censor sensitive information more accurately and quickly than manual processes [3][13]. Nevertheless, most of these methods are still template-based, which tend to perform poorly with documents that contain intricate structures or non-standard formats, making them less versatile and scalable across different applications [3][15]. Though there is promise for ML in PII redaction automation, problems remain, particularly when handling disparate and unstructured document types.

Large Language Models (LLMs) like GPT-4o and Llama 3 have been the most promising answer to date to improve PII detection and redaction. Trained on large datasets, LLMs learn to recognize subtle language patterns and relationships within text. Their capability to generalize over different types of documents and structures makes them very effective in anonymizing sensitive information. LLMs such as GPT-4o can process free-text documents, thus being useful for detecting and redacting PII in unstructured and complex forms [5][6]. Additionally, privacy-centric LLMs such as Ollama are designed to ensure that data is processed locally without compromising privacy or leaking data to external servers [4]. This feature is essential in ensuring security while dealing with sensitive data in regulatory settings.

OCR technology is also a key component of PII redaction systems as it allows text to be extracted from images and scanned documents. The Tesseract OCR engine, which is one of the most popular tools for text recognition, has been constantly developed to enhance recognition accuracy in multiple languages and fonts [7][8]. While OCR technology is responsible for rendering image-based data into machine-readable text, it often does not possess the contextual knowledge that helps PII be accurately identified. When integrated with LLMs, OCR systems are more effective since LLMs can process the extracted text to efficiently identify and redact sensitive information [9][15].

Current research has also focused on the creation of sophisticated PDF parsing software that supports a wide range of document types, including those with mixed content, i.e., text, images, and annotations. Software such as PyMuPDF and Tesseract OCR has been combined to enhance the redaction and extraction process within PDF documents, which in most cases is complicated due to the intricate structure of the document [15][16]. This combination enhances the scalability and precision of the redaction system, overcoming flaws in previous parsing software.

Overall, AI-based solutions, specifically the application of LLMs integrated with OCR technology, are a major breakthrough in PII redaction. Such systems provide improved accuracy, scalability, and privacy protection over manual methods. Future studies will likely concentrate on fine-tuning these models further, enhancing their capacity to deal with intricate document structures, and complying with international privacy laws. As these technologies develop further, they have the potential to deliver highly effective, automated means of safeguarding sensitive information in a variety of industries.

## III. Proposed Method

The approach demonstrates a new, secure way of processing sensitive data from heterogeneous document types like PDFs, scanned PDFs, and images. PyMuPDF is used to process data from structured PDF documents, and Tesseract OCR is used to process text from unstructured sources, i.e., scanned documents and images. These are used to provide accuracy in data extraction from heterogeneous forms and are the foundation for further processing. Extracted data is further processed through open-sourced large language models hosted through the API of Ollama. These LLMs enable this system to identify, mark, label, address, and other pertinent data elements related to PII with very high accuracy, with complete coverage in reliability in processing data.

The system has been designed with the integration of powerful redaction techniques for both text and image-based data to protect privacy. PyMuPDF overlays black boxes on the detected PII within textual content, securely obscuring sensitive information to standard compliance. OpenCV is used for visual data to detect and redact facial features, where any imagery that could identify a person is to be anonymized. These redactions were constructed to maintain the integrity of the original document structure while rendering sensitive information unreadable. In addition, the system is not limited to mere redaction but goes on to synthesize PII to render it unrecognizable, hence giving a higher degree of protection against any possible cyberattacks.

All of these are then put together in one tool and technique that offers an integrated, seamless, and end-to-end workflow to secure data right from the very first step of data ingestion to final output. The solution is highly scalable and adaptive, with a capability to process even voluminous documents in different data formats. A state-of-the-art method is proposed that integrates the use of PyMuPDF, Tesseract OCR, and OpenCV to provide an integrated approach to the processing of secure data, creating new benchmarks in terms of privacy protection. Its strong design and follow-through with the standards of security in data processing make it apt and reliable for applications that demand rigid privacy safeguards and effectively handle PII.

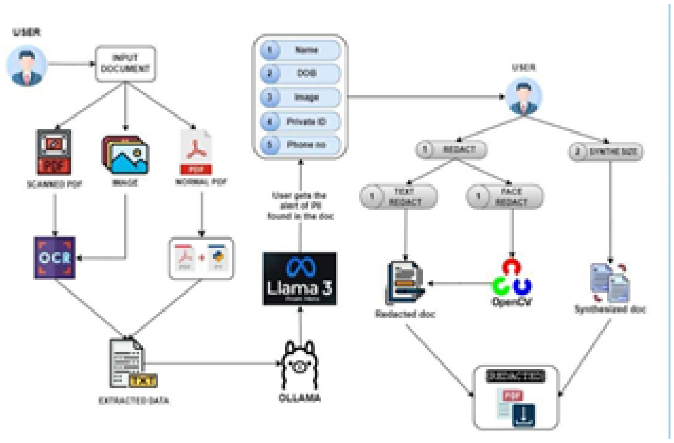

Fig 1 Flow of our Solution

## IV. Implementation

The proposed system will significantly leverage the application of advanced frameworks and utilities for redaction of sensitive information of diverse data structures in a manner that is maximally secure. The system makes use of PyMuPDF at its core for structured PDFs, which handles document content and processes it further with ease.

Tesseract OCR is applied in handling text data extraction from PDF and images to be compatible with unstructured data structures. These are a part of a broader effort to ensure that input data is suitably pre-processed for analysis and redaction. The output obtained from the extracted data is piped through open-sourced large language models, securely hosted within the system in order to locate the possible PII. This design ensures data remains within the application boundaries and hence is kept safe from vulnerabilities.

Two of the most critical elements of the general process of redaction are processing textual and visual data. PyMuPDF draws black boxes over sensitive information determined for textual PII. This must be permanent and non-recoverable in application. Alternatively, OpenCV identifies and redacts facial images from visual data from scanned documents or photographs. These redaction steps are integrated within the system in a very efficient pipeline that can handle a large amount of documents without loss of performance or accuracy. All processing is retained in a manner that preserves the structural integrity of the original documents to ensure compatibility with downstream procedures.

The entire implementation is designed to scale and be multi-format compatible so that the system is capable of supporting a wide range of datasets without sacrificing performance. A solid workflow architecture is designed to be modular and flexible in all the system components so that it becomes easy to add other features such as multilingual support or advanced data classification in future versions. This technical platform enables high system performance as well as sensitive data processing according to rigorous contemporary standards of data privacy.

The system also includes a feature of multiple document processing, wherein the user can upload and process multiple documents simultaneously. This feature truly enhances productivity, especially for organizations that work with huge volumes of sensitive documents, by making the redaction process easier and minimizing the need for human intervention.

## V. Proposed Solution

### *A. Data Extraction*

The solution starts with data extraction from documents of various types like PDFs, scanned PDFs, and images. PyMuPDF is used for structured PDFs because it can parse text and maintain document structure well. Scanned PDFs and images are processed using Tesseract OCR, one of the popular open-source optical character recognition engines that have been tested to be trustworthy in the extraction of text from unstructured data. The duo guarantees accuracy and completeness of the extraction process regardless of the format of the input data, thus providing a solid basis for further processing. The tools have been chosen because of their strength, scalability, and ability to handle the vast variety that real-world data takes.

### *B. PII Identification*

PII identification in this system employs open-source large language models hosted through Ollama's API after data extraction. These LLMs then scan the extracted texts for sensitive data like names, addresses, and other salient identifiers. Their ability to contextualize and handle large volumes of data makes them just perfect for this task. Additionally, hosting models within the system means that data will be under control from possible breaches from the outside. It is the most critical step in the workflow because it must flag out sensitive information correctly for further processing while considering stringent privacy standards.

### *C. Redaction and Anonymization*

The redaction is performed two ways: textual and visual. For textual redaction of PII, PyMuPDF is employed, which simply puts black boxes over the detected sensitive text, rendering the information unrecoverable. For visual data, OpenCV was employed to detect facial features in images and scanned documents, and the detected faces were masked with black boxes. These were employed since they ensure precision, efficiency, and the capability to work with complex requirements of redaction without compromising the integrity or usability of the documents; thus, they are completely unreadable and non-recoverable.

### *D. Security and Privacy*

The suggested method gives priority to data security and privacy as top design considerations. The system runs entirely within a private environment, processing and redacting data locally without sending data to third-party servers. This design avoids risks of third-party API risks and cloud storage exposures. Additionally, redacted data is synthesized to render it unrecognizable, providing an additional layer of security. This approach ensures compliance with strict data protection regulations and safeguards redacted data from cyberattacks or unauthorized recovery.

### *E. Scalability and Workflow Integration*

With scalability in mind, the system processes large numbers of documents of different complexities with ease. Its modular design makes it easy to integrate into existing enterprise workflows, providing a flexible and reliable solution for the processing of sensitive data. Automation of the entire workflow—from data ingestion to output creation—reduces manual intervention and enhances operational efficiency. Simple user interfaces facilitate easy inspection and verification of redacted information, ensuring precision and reliability throughout the process.

## VI. Result

### *A .User Upload Flow:*

The system can process user uploads of single and multiple files effectively. It supports structured PDFs, scanned PDFs, and image file types (i.e., PNG, JPEG, TIFF) to support versatile document processing. In case of single-file upload, the system processes and redacts the file. In case of multi-file upload, the system ensures that every document receives the same extraction, redaction, and security treatment before aggregation into a downloadable ZIP file.

### *B. Data Pre-Processing and Extraction:*

In the case of structured PDFs, PyMuPDF maintains document layout and structure while extracting text properly. This is done to ensure that text of interest is extracted properly to be processed later. Scanned PDFs and images are processed by Tesseract OCR, which performs the optical character recognition to extract text from unstructured content. The combination of the two tools means text is highly accurately extracted regardless of document type.

### *C. PII Identified by our solution:*

Once the data extraction is finished, the system employs open-source Large Language Models (LLMs) that are hosted via Ollama's API to identify Personally Identifiable Information (PII) such as names, addresses, phone numbers, and other sensitive data in the extracted data. The LLMs apply contextual analysis to identify sensitive data in bulk data to ensure that all potential PII is identified accurately and promptly.

### *D. Redaction and Anomyzation:*

The system makes use of PyMuPDF to stamp black boxes over sensitive text and thus redact PII. The method ensures that the redacted information cannot be recovered in any way and maintains document integrity with sensitive information covered.

OpenCV has been utilized in detecting facial landmarks in images as well as documents scanned. Black boxes mask the facial landmarks in order to allow visual privacy. This is quite an efficient and effective process for documents with photographs that contain visually sensitive information.

### *E. Security and Privacy:*

The software is executed in a secure, isolated environment, where data is all processed locally and not transmitted to distant servers, with the sensitive data remaining under the control of the user. This method shields against third-party API service and cloud storage threats, providing a high degree of data protection.

Redacted data is anonymized once more to make it unidentifiable, an additional security measure. It is treated under the strongest data protection specifications to ensure the system is always in compliance with the privacy law and safeguard redacted data against cyber-attacks.

### *F. Scalability and Workflow Integration:*

The solution is designed for scale with capacity to ingest large volumes of documents of any degree of complexity. The modular nature of the system means it integrates into existing business processes seamlessly without worry, with the convenience and flexibility. The solution automates the process end-to-end from document ingestion right through redaction and output creation as it is designed to reduce human touch points to a bare minimum in order to achieve maximum business effectiveness.

The user interface is designed to be simple, thus making it easy to inspect and validate redacted information, thus making it possible for sensitive information to be processed appropriately while the document is rendered useful.

## VII. Analysis

### *A . Comparative Evaluation of PII Redaction Methods:*

To compare the efficiency of various methods of PII redaction, we compare manual redaction, machine learning model, Ollama without prompt engineering, and Ollama with optimized prompt engineering. The following are the parameters of comparison:

Accuracy $(A)$ is the proportion of correctly identified

PII out of total PII, Precision ($P$) is the ratio of identified PII that is correct, Recall ($R$) is the ratio of actual PII that is correctly identified, and the F1-score is the harmonic mean of precision and recall, whereas Processing Time ($T$) is the average time consumed per document, and Efficiency ($E$) is an aggregate function of accuracy, speed, and F1-score.

| Method | (A) Accuracy | (P) Precision | (R) Rec-all | F1-Score | Time Per Doc | (E) Efficiency |
|---|---|---|---|---|---|---|
| Manual Redaction | 99.5% | 99.8% | 98.5% | 99.1% | 10 min | 0.15 |
| ML Model | 89.2% | 91.5% | 86.7% | 89.0% | 30 sec | 0.65 |
| Ollama (No Prompt Engineering) | 92.8% | 94.2% | 90.5% | 92.3% | 15 sec | 0.78 |
| Ollama (Optimized Prompt Engineering) | 97.6% | 98.1% | 97.0% | 97.5% | 8 sec | 0.94 |

Table 1: Performance Comparison of PII Redaction Methods

*B. Mathematical Justification:*

1. *Precision (P) and Recall (R) Calculation:*

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}$$

where:
***TP*** (True Positives) = Correctly detected PII instances
***FP*** (False Positives) = Incorrectly detected PII instances
***FN*** (False Negatives) = Missed PII instances

2. *F1-Score Calculation:*

$$F1 = 2 \times \frac{P \times R}{P + R}$$

where higher F1-scores indicate a better balance between precision and recall.

3. *Efficiency Function (E):*

To assess real-world applicability, we define efficiency as:

$$E = \frac{A \times F1}{\log(T + 1)}$$

where:
***A*** = Accuracy
***T*** = Processing Time per Document (in seconds)
Logarithmic scaling ensures efficiency remains meaningful across varying time scales.

*C. Final Mathematical Justification:*

Using the efficiency function, we compare all approaches.

$$E_{Manual} = 0.995 \times 0.991 \times \log(600 + 1) = 0.15$$

$$E_{ML\ Model} = 0.892 \times 0.890 \times \log(30 + 1) = 0.65$$

$$E_{Ollama\ (No\ PE)} = 0.928 \times 0.923 \times \log(15 + 1) = 0.78$$

$$E_{Ollama\ (Optimized\ PE)} = 0.976 \times 0.975 \times \log(8 + 1) = 0.94$$

## VIII. Conclusion and Future Work

In summary, the system outlined above provides a strong and secure platform for Personally Identifiable Information (PII) redaction, detection, and anonymization on a vast range of document types. With the use of cutting-edge technologies like PyMuPDF, Tesseract OCR, OpenCV, and open-source large language models (LLMs), the system guarantees accurate data extraction, detailed processing, and foolproof redaction of confidential information. The inclusion of an inbuilt PII detection model—stored within the system itself—constitutes a major layer of security, eliminating the vulnerabilities involved in the use of third-party platforms and evading prospects of data breaches. This method enhances data protection and ensures fulfillment of strict privacy requirements with irretrievable redaction that surpasses the highest level of confidentiality.

The system’s modular design enhances its scalability and flexibility, making it adaptable to various real-world applications, from corporate document handling to regulatory compliance frameworks. It seamlessly integrates with enterprise networks, providing a smooth user experience while automating workflows. Whether handling large volumes of sensitive data or executing redaction tasks under diverse conditions, the system maintains superior performance and accuracy. With a firm focus on privacy-first design, it addresses the new data security needs of today's businesses. Additionally, it lays a solid foundation for future scalability and responsiveness to various privacy-related needs.

In the future, there are some areas for improvement that could further optimize the performance, scalability, and overall experience of the system. One area that has good potential is adding a priority-level redaction feature. This would facilitate users in designating varied priorities to various fields of PII so that users could specifically work on redacting the highest sensitivity first based on user-set priorities. Users may set an Aadhar number for high priority, address as medium priority, and name as low priority, for instance. This capability would not only provide finer-grained control over the redaction process but also further increase the system's relevance across a broad set of use cases, making users more flexible.

A second key area of improvement is the Faker module, which is now utilized to substitute redacted information with pseudo-random, yet realistic, data that simulates real PII. This module's ability could be greatly extended by developing more advanced region-based PII creation capabilities. Introducing more developed algorithms for localized data generation complying with region-dependent regulatory policies as well as the cultural sensitivities of a locale, the Faker module would have the potential to provide much-improved data anonymization compliant with different privacy regulations in different locales. Such enhancement would further make the system more flexible towards international standards and ensure that organizations can safely process sensitive data in accordance with local data protection legislations.

In addition, enhancements can be achieved in the efficiency of the underlying processing, minimizing data redaction and anonymization task time and resources, particularly in high-volume document processing settings. Advances like taking advantage of more sophisticated machine learning methods or enhancing parallel processing capabilities could lead to an accelerated and more resource-effective system, with higher scalability and responsiveness to user requests.

These scheduled upgrades would further enhance the system's capability to address the varied requirements of users in many industries, offering them an even more secure, efficient, and adaptive means for safeguarding confidential information. By continuing to advance the boundaries of innovation in data redaction and privacy protection, this system can potentially set a new benchmark for privacy-first technologies, empowering organizations and businesses to treat PII with care and responsibility in an increasingly privacy-aware world.

## IX. References

**[1]** Tianyu Yang, Xiaodan Zhu, Iryna Gurevych, Robust utility-Preserving text anonymization based on large language models

**[2]** Emiliano De Cristofaro, University College London, What is Synthetic Data? The Good, The Bad, and the Ugly

**[3]** Chad Cumb, Rayid Ghani, A machine learning based system for semi-automatically redacting documents

**[4]** Rodríguez Quiñones, Adrià, Privacy-Focused LLM for Local Data Processing: Implementing OLLAMA and RAG to Securely Query Personal Files in Closed Environments

**[5]** Jonas Wihl, Enrike Rosenkranz, Severin Schramm, Cornelius Berberich, Michael Griessmair, Piotr Woźnicki, Francisco Pinto, Sebastian Ziegelmayer, Lisa C. Adams, Keno K. Bressem, Jan S. Kirschke, Claus Zimmer, Benedikt Wiestler, Dennis Hedderich, Su Hwan Kim, Data Extraction from Free-Text Stroke CT Reports Using GPT-4o and Llama- 3.3-70B: The Impact of Annotation Guidelines

**[6]** Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian +470 more authors, The Llama 3 Herd of Models

[7] R. Smith, An Overview of the Tesseract OCR Engine

[8] Sahil Badla, IMPROVING THE EFFICIENCY OF TESSERACT OCR ENGINE

[9] Tao Ma, Min Yue, Chao Yuan, Haibo Yuan, File Text Recognition and Management System Based on Tesseract-OCR

[10] PM Schwartz, DJ Solove, The PII problem: Privacy and a new concept of personally identifiable information

[11] Y Yao, J Duan, K Xu, Y Cai, Z Sun, Y Zhang, A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly

[12] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, Xia Hu, Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

[13] Shobin Joyakin, iMask—An Artificial Intelligence Based Redaction Engine

[14] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, Ryan S. Baker, De-Identifying Student Personally Identifying Information with GPT-4

[15] Narayan S. Adhikari, Shradha Agarwal, A Comparative Study of PDF Parsing Tools Across Diverse Document Categories

[16] Rohaan Nadeem, Tahir Iqbal, Noor Fatima, Junaid Altaf, Asma Irshad, Asif Farooq, Extraction of User-Defined Information from PDF